an audio pickup device for generating audio signals representative of sound from an audio source; and

a multimodal integration architecture system for processing said image signals and said audio signals to determine a direction of the audio source relative to a reference point.

2. (original) The video conferencing system of claim 1 wherein said multimodal integration architecture system further comprises:

an audio source localization system;

a computer vision person detection system; and

a multimodal speaker detection system.

3. (original) The video conferencing system of claim 2, further comprising an integrated housing for an integrated video conferencing system incorporating the image pickup device, the audio pickup device, and the multimodal integration architecture system.

4. (original) The video conferencing system of claim 3, wherein the integrated housing is sized for being portable.

5. (original) The video conferencing system of claim 2, further comprising an electronic pan tilt zoom system for electronically manipulating the image signals to effectively provide at least one of variable pan, tilt, and zoom functions.

6. (original) The video conferencing system of claim 5, wherein the image pickup device is a stationary camera.

7. (original) The video conferencing system of claim 5, wherein the multimodal integrated architecture system provides control signals to the electronic pan tilt zoom system.

8. (original) The video conferencing system of claim 7, wherein the audio source moves relative to the reference point, the audio source localization system detects the movement of the audio source, and, in response to the movement, the audio source localization system causes a change in the field of view of the image pickup device.

9. (original) The video conferencing system of claim 5, wherein the audio pickup device is comprised of an array of two microphones.

10. (currently amended) A method comprising the steps of:

generating, at a stationary ~~an~~ image pickup device, remaining motionless during operation, image signals representative of an image;

generating, at an audio pickup device, audio signals representative of sound from an audio source;

processing the image signals and the audio signals to determine a direction of the audio source relative to a reference point;

manipulating the image signals to produce refined image signals; and

outputting said refined image signals.

11. (original) The method of claim 10 further comprising the steps of:

applying said audio signals to an audio source localization system;

applying said image signals to a computer vision person detection system;

processing said audio signals and said image signals with a multimodal speaker detection system;

generating control signals based on the determined direction of the audio source;

applying the control signals to an electronic pan tilt zoom system to mimic the effect of at least one function of a movable camera, said function selected from the group consisting panning, tilting, and zooming said movable camera; and

providing an output from said electronic pan tilt zoom system.

12. (original) The method of claim 10, further comprising electronically varying a field of view of the image pickup device in response to the control signals.

13. (original) The method of claim 10, wherein processing the audio signals includes determining an audio based direction of the audio source based on the audio signals.

14. (original) The method of claim 12, wherein the audio source moves relative to a reference point, and wherein processing the audio signals further includes:

detecting the movement of the audio source; and

causing electronically, in response to the movement, an increase in the field of view of the image pickup device.

15. (original) The method of claim 12, further comprising the step of supplying control signals, based on the audio based direction, for electronically panning, tilting, or zooming said image pickup device.

16. (currently amended) A video conferencing system comprising:

two microphones for generating audio signals representative of sound from a speaker;

a stationary video camera, remaining motionless during operation, for generating video signals representative of a video image;

an electronic pan tilt zoom system for manipulating video images to produce the visual effects of panning, tilting, and/or zooming;

a processor for processing the video signals and the audio signals to determine a direction of a speaker relative to a reference point and supplying control signals to the electronic pan tilt zoom system for producing images that include the speaker in the field of view of the camera, the

control signals being generated based on the determined direction of the speaker; and

a transmitter for transmitting audio and video signals for video conferencing.